DOCUMENT RESUME

ED 244 971 TM 840 262

AUTHOR Bliss, Leonard B.
TITLE Basic Skills Achievement in the Caribbean: A Research Model.
PUB DATE Apr 84
NOTE 29p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Tests; Basic Skills; Elementary Secondary Education; Item Analysis; *Sampling; *Standardized Tests; Test Reliability; *Test Validity; Timed Tests
IDENTIFIERS Item Discrimination (Tests); *Stanford Achievement Tests; Stanford Test of Academic Skills; *Virgin Islands

ABSTRACT

A model for the validation of standardized tests of academic achievement upon populations not represented in the samples used to standardize the tests is presented, and the results of a field testing of the model are described. The 1973 editions of the Stanford Achievement Test and the Test of Academic Skills were administered to a sample of predominantly West Indian students in the public schools of the Virgin Islands of the United States. Analysis indicated characteristics similar to those obtained from the contintental United States standardization sample in terms of reliability, content validity, and item discrimination indices. Item analysis revealed differences between the local and standardization samples based on the cognitive complexity of items on all subtests. There were also indications of effects of local dialects on responses to language subtests. Finally, the data indicated that most students were unable to complete the reading comprehension subtests in the standard time allotted. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED244971

Basic Skills Achievement in the Caribbean:

A Research Model

Leonard B. Bliss
Caribbean Research Institute
College of the Virgin Islands

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Leonard Bliss

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 840262

Presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA, April 23-27, 1984

Abstract

A model for the validation of standardized tests of academic achievement upon populations not represented in the samples used to standardize the tests is presented and the results of a field testing of the model are described. The 1973 edition of the Stanford Achievement Test and the Test of Academic Skills was administered to a sample of predominantly West Indian students in the public schools of the Virgin Islands of the United States. Analysis indicated characteristics similar to those obtained from the contintental United States Standardization sample in terms of reliability, content validity, and item discrimination indices. Item analysis revealed differences between the local and standardization samples based on the cognitive complexity of items on all subtests. There were also indications of effects of local dialects on responses to language subtests. Finally, the data indicated that most students were unable to complete the reading comprehension subtests in the standard time allotted.

The widespread use of standardized achievement testing in the English speaking Caribbean has posed a series of problems for educators in this area of the world. Not the least of these problems involves the reliability of scores obtained from students to whom these tests are administered.

During the colonial period, curriculum was imported, as a more or less complete package, from the mother country, complete with form examinations which were designed and standardized overseas. Local school people had little or no autonomy in terms of curriculum or evaluation procedures. With the coming of independence and the emergence of these former colonies into nationhood, this control has disappeared and national ministries and departments of education now play the major role in determining the curriculum, including evaluation procedures, which will be used in their schools. While most of these emerging nations till hold strong emotional and cultural ties to their former mother countries, there are strong pressures for their educational systems to move toward more independent, locally relevent curricula with the accountability this type of movement would dictate. Valid and reliable tests of achievement are a necessary component of this accountability.

The use of standardized, commercially published achievement tests offers much to recommend them as instruments toward meeting the goal of high standard evaluation. The items on these tests tend to be technically superior to those found on informal tests. They have gone through a series of trials and revisions and have met standards of clarity and precision that tend to be high and well defined. In addition, much is known about typical

performances of students in a particular population when they are administered these tests. Also, test publishers go to great efforts to determine the curriculum used in the target population schools and to include items which constitute a representative sample of the cognitive objectives in the curriculum at these institutions. Finally, machine scoring is usually available for these tests and results may be reported out either in criterion referenced form or norm referenced based on a reasonably well defined population.

It is in these latter two areas that English speaking Caribbean jurisdictions find major difficulties. Published standardized tests of achievement are standardized using non-Caribbean peoples. This is not supprising considering the small populaton of the area and the resulting small markets for such tests. The costs of producing high quality tests are extremely high and publishers must look toward large markets when planning new tests and revisions of prexisting examinations. As a result of this, it is hard for people involved in decision making at Caribbean ministries and departments of education to be confident that a given test or series of tests evaluates a representative sample of the objectives in their curricula, i.e. that the test is content valid.

Additionally, while the items used on standardized achievement tests seem to function well for examinees in the population from which the standardization sample was drawn, there may be some justifiable concern about whether or not these items will function in a similar manner when administered to Caribbean

students since these students were not part of the population from which the standardization samples were drawn (generally populations of students in the continental United States, the United Kingdom, or Canada). Given the possibility of cultural factors affecting test taking performance, Caribbean decision makers might well be justified in being hesitant to accept the results of standardization procedures reported by the publishers of these tests.

An obvious solution this this dilemma is for each ministry and department of education to develop and standardize a set of achievement tests appropriate for the content of its curriculum and the test taking characteristics of its students, keeping in mind the need to create alternative forms of these tests and the need to update them, periodically. While this may initially appear to be a workable procedure, the costs in time and money are probably beyond the resources of most of these educational systems. In addition, the technical expertise required to carry out such an effort would most likely be unavailable locally and the importing of persons from the outside prohibitively expensive and undersirable in other ways.

A second solution would be convincing commercial test publishers to produce achievement tests which were content valid for local curricula and standardized on the local populaton of students. It seems unlikely that this effort would bear any fruit based on geogaphical consideratons and the relatively small market that would be available for such tests.

The purpose of this paper is to suggest a third alternative to solving the problem of obtaining valid and reliable

standardized tests of academic achievement for use in the English speaking Caribbean and to report on the procedures and the results of using this model in a field setting. The model is based on the assumption that while there may have been some changes in curriculum since political independence, most Caribbean states retain much of the educational structure that existed during the colonial period due to strong emotional and cultural ties to the former mother country and the usual strong conservative predisposition of educational systems in most democratic countries. These are not unreasonable assumptions and the latter is supported by the facts that most high level school officials involved in decision making capacities at ministries and departments of education in the Caribbean received all or part of their training in American, British, or Canadian colleges and universities, and that most of the texts used in the English seaking Caribbean are published in these three countries. The latter is particularly important in light of the fact that, in most classrooms, the content of the curriculum is determined primarily from the content of the text book or books being used.

Under these assumptions, the model proposes that published standardized tests used and standardized in school systems similar to those in question be examined, first to determine the content validity of these tests given the curriculum in the local school system, and then to establish the reliability of scores and the test taking behaviors of a representative sample of local students in order to determine the appropriateness of the chosen test. Finally, if the test appears to function well for the

population of local students, adjustments to the test items and/or test taking instructions can be made based on the results of the local validation study before the test is placed in use on a system wide basis.

The Testing Site

Improving basic skills achievement was a concern of the Department of Education of the Virgin Islands of the United States when it approached the College of the Virgin Islands to provide aid in improving instruction in these areas. In an effort to provide this service, the Caribbean Research Institute, the college's research arm, worked with a task force composed of representatives from the Department and the Institute to determine a course of action.

It became clear after the first few task force meetings that the development of any strategy designed to improve basic skills achievement needed to start off with a fairly detailed description of current achievement levels of students in territorial schools. This information was not available since the school system had no program of standardized testing in place. Various published achievement tests from the continental United States were administered from time to time at the discretion of building principals, but the test used and the time of testing were at the whim of these administrators and the records kept of these results were rather haphazard. The Iowa Test of Basic Skills was administered, system wide, to sixth graders, but some building administrators often refused to have these tests administered in their schools and there were years, due to fiscal

constraints, when the tests were not given, at all. Even the scores obtained were of little use since they were reported out in norm referenced form based on U.S. national and local norms and provided no information as to the particular skills students posessed or lacked. Finally, none of the tests used had been validated using local students and there was a strong feeling that cultural and curricular differences between mainland U.S. and U.S. Virgin Islands students and schools rendered the reliability and validity of these scores questionable. There were no standardized tests of academic achievement administered on the secondary level anywhere in the territory's public schools.

The task force decided to test the appropriateness of an achievement test of basic skills devised and standardized in the continental United States when it was administered to students in the U.S. Virgin Islands public schools. The Virgin Islands of the United States is an unincorporated territory of the U.S. comprising some 50 islands and cays in the Caribbean Sea. The two largest islands, St. Thomas and St. Croix, are separated by 40 miles. The island of St. John lies three miles east of St. Thomas. The total land mass of the territory is 132 square miles. Only the three largest islands have a sizable permanent population, estimated at about 120,000. This is augmented by a transient population of almost one million tourists each year. Of the permanent population, approximately 80% are of West Indian heritage, either having been born in the U.S. Virgin Islands or having immigrated from other islands in the Lesser Antilles. St.

Croix has a significant Hispanic population, originally from Puerto Rico and its smaller islands of Vieques and Calebra. The official language of the territory is English with many persons speaking a patois derived from English, Dutch, and French at home and in informal circumstances.

The K-12 populaton of the public schools is approximately 25,000 with education being compulsory from age six to sixteen. Standard English is the language of instruction. The population of students attending USVI public school is primarily West Indian and Hispanic. Approximately 94% of students attending are entitled to free lunch under the U.S. Department of Agriculture's school lunch program. The vast majority of residents from the continental U.S. and other middle socioeconomic status families send their children to one of the many private day schools in the territory.

Although it is separated from the U.S. mainland by 1100 miles of ocean, the USVI is hardly isolated. The three local television stations broadcast network television (including PBS) and television stations from Atlanta, Chicago, and New York are available on cable television. New York and Miami newspapers are available on a same day basis and U.S. magazines are available on a regular basis. Nor is the school system curricularly isolated. Most basic skills curriculum is imported, intact, from the continental United States. Reading is taught using the Ginn 720 series and mathematics usingthe series published by Silver-Burdette, Co., for instance. English grammar is taught using the time honored series by Warriner and literature texts published by mainland U.S. publishers are used in both elementary and

secondary schools.

Teachers in the public school system tend to have been trained primarily at the College of the Virgin Islands, the territory's public land grant college, or at mainland U.S. colleges and schools of education. The former provides a standard U.S. college curriculum with a traditional program of teacher education.

Given these similarities in curriculum and teacher preparation, and the degree of communication with the continental U.S., the task force agreed to choose and attempt to validate a standardized test of basic skills which had been published in the mainland U.S. and used in mainland U.S. schools.

The Instrument

In choosing a test battery to be validated on U.S. Virgin Islands students, the following criteria were used:

1) The test needed to be technically sound in terms of reliability and item descrimination, at least for the population which it had been standardized on.

2) The test needed to be content valid for U.S. Virgin Islands public school students. That is, the test needed to contain items which tested a representative sample of the content and behaviors actually taught at various levels in the USVI public schools.

3) The test publishers needed to include a detailed description of the objectives tested while providing an item by objective keying procedure.

4) Scores which indicated students' performances relative to each objective needed to be available. That is, criterion referenced scoring was a requirement.

The 1973 version of the Stanford Achievement Test and the Test of Academic Skills was chosen as the test battery which best appeared to meet the above criteria.

Validation Procedure

Sampling The June 1, 1980 enrollment in the public schools in the Virgin Islands of the United States was 25,426 according to the statistics issued by the USVI Department of Education. It was clear that testing this number of students was economically unfeasible. The preferred alternative would have been to generate a random sample of students in grades K-12 to be tested, but it was equally clear that this would have produced an intolerable disruption of classroom activities. Therefore, in an attempt to obtain a representative sample of students, cluster sampling was used with the clusters being defined as classes. The number of classes to be selected for the sample from each grade in each of the St. Thomas/St. John and the St. Croix school districts was determined by calculating the proportion of the total K-12 student population in each grade in each district and assuming a class size of 30 in the elementary schools and 27 in the secondary schools.

Selecting whole classes presented an additional difficulty. The small number of classes selected in each grade might have made obtaining a representative sample of students more difficult. This is due to the fact that while classes in a given

elementary school may be heterogeneous, the schools themselves are not. This is because elementary schools in the U.S. Virgin Islands are essentially neighborhood schools. Virgin Islands neighborhoods tend to be homogeneous in terms of socioeconomic status of residents. To overcome this problem, it was decided to increase the number of classes tested in a given grade (thereby increasing the number of schools within the territory from which these classes came) without increasing the total number of students tested by testing at alternate grades. This seemed acceptable since many of the objectives tested by the Stanford Achievement Test carry across adjacent levels of the test and there was no reason to suspect that the patterns of academic achievement of students in even numbered grades were different from those in odd numbered grades. Classes in grades 2, 4, 6, 8, 10, and 12 were given the test. Bliss (1982) describes this procedure in detail and comments on the effects of sampling classes rather than individual students.

Testing Procedures Testing was done at the grade level recommended by the test publisher. This was done primarily to insure the content validity of the examinations. Tests were administered by classroom teachers or guidance counselors, at the discretion of building administrators. Each person who was to administer tests attended a two hour training session at either the College of the Virgin Islands St. Thomas or St. Croix campuses. During this time the purpose of the testing was explained, the test and instruction manual were reviewed, a testing schedule was distributed and reviewed, and testing

materials were distributed. These included a practice test to be given to students in grades 2, 4, and 6 the day prior to the first day of testing in order to give these students experience in reading and answering items on this type of test. Answer sheets were sent off island to be machine scored.

Content validity The content validity of the various levels of the Stanford Achievement Test was determined using the following strategies:

1) Collection of written curriculum guides used in the public schools. The objectives explicitly stated or implicitly inferred in these documents were compared with the lists of objectives tested provided by the test publisher.

2) Text books used in the teaching of basic skills subject matter were collected from selected schools. Stated and implicit objectives in these texts were compared with the test publisher's objectives.

3) The test objectives were shown to elementary and secondary subject area supervisors who were asked to determine the degree of match between these objectives and what was taught at the indicated grade levels.

4) Selected building principals in St. Thomas were asked to review the objectives of the test and give their opinions concerning the degree of match between these objectives and the objectives taught toward in their schools.

5) Teachers who administered the tests in their classrooms were asked to review the test publisher's objectives and to determine the degree of match between these objectives

and the basic skills they expected their students to have obtained.

Using these techniques, the researchers were satisfied that the test did, indeed, test a sample of objectives that was consistent with the objectives used in teaching in the public schools of the Virgin Islands of the United States.

Reliability Kuder-Richardson 20 estimates of internal consistency were calculated for each test of each battery for the entire USVI sample and the subsamples from each of the two school districts. It was noted that, in most cases, the variances of the raw scores obtained by the USVI sample of students were considerably lower than those reported for the standardization group. This is not an uncommon phenomenon and is commonly found when testing samples drawn from populations composed largely of persons from low socioeconomic status homes. Since the reliability of a test is partially dependent on the heterogeneity of the scores obtained (the greater the spread of scores, the higher the reliability), the local scores were adjusted for homogeneity using a procedure described by Allen and Yen (1979). See Bliss (1982) for details of this procedure.

There needed to be a criterion for making decisions regarding the acceptability of the adjusted reliability estimates. The Stanford Achievement Test is considered to have more than acceptable reliability when administered to the population of examinees upon which it was standardized (i.e. continental U.S. students). Among the indications of this are numerous reviews of the test in the literature (Kasdon, 1974; Lehmann, 1975; Chase,

1978; Ebel, 1978; Thorndike, 1978) and the fact that it is widely used in the schools. However, the literature is replete with studies which indicate that standardized tests of academic achievement tend to produce less reliable scores when administered to students from low socioeconomic status homes and to those who are culturally different from the majority of those on whom the test was normed (see reviews and discussions in Anastasi, 1958; Tyler, 1956; and Deutsch, 1960). Therefore, if the reliability estimates obtained from a sample of U.S. Virgin Islands students who took the Stanford Achievement Test are at least equal to the reliability estimates obtained from the standardization samples, it is reasonable to conclude that the test scores are reliable indicators of academic achievement for these students.

For each adjusted reliability estimate obtained from the USVI sample, a reliability difference was found by subtracting the standardization group's reliability estimate from the local group reliability estimate. The median reliability difference across all tests for all grades was -.002 with a range from -.06 to +.02 with the distribution somewhat negatively skewed. When Z transformations were used to normalize the distribution of the reliability estimates, t-tests revealed two subjects out of the total 36 examined where the local sample reliability estimates were significantly lower than those of the standardization group at the p=.05 level (see Bliss, 1982). This is approximately the number that would be expected by chance. The standard errors of measurement (which are not affected by the variances of the scores) were treated in a similar manner and it was found that

there were only three out of 36 standard error estimates which were significantly higher than those obtained from the standardization sample.

Item discrimination The item discrimination index indicates the degree to which responses on one item are related to responses on other items of the test. The statistic indicates whether a person who does well on the test as a whole (that is, a person who is presumably high on the trait being measured) is more likely to get a particular item correct than a person who does poorly on the test as a whole. In other words, the item discrimination index indicates whether an item discriminates between those who do well and do poorly on the test as a whole. Taking the item difficulty and the item discrimination index into consideration, the developers of tests desire to construct tests which discriminate well among examinees with varying levels of a trait.

The item discrimination index is calculated by the formula $d=(U-L)/N$ where

U= the number of examinees who have total test scores in the upper range of total test scores and who also have the item correct.

L= the number of examinees who have total test scores in the lower range of total test scores and who also have the item correct.

N= the number of examinees in the upper or lower range of the test scores.

By definition, d is the difference between the proportion of high scoring examinees who got the item correct and the proportion of low scoring examinees who got the item correct. The upper and lower ranges generally are defined as the upper and lower 10% to 33% of the sample, with examinees ordered on the basis of their total test score. When total scores are normally distributed, using the upper and lower 27% produces the best estimate of d (Kelly, 1939). If the distribution of total test scores is flatter than the normal curve, the optimum percentage is larger and approaches 33%. However, Allen and Yen (1979) found that, for most applications, any percentage between 25 and 33 will yield similar estimates of d. In this study, 27% was used as the upper and lower percentages because examination of selected distributions of actual test scores revealed nearly normal distributions.

The theoretical range of d is between -1 and +1. However, maximum discrimination is likely to occur when the difficulty index equals .50. When p=.50 the variance in item scores, which is p(1-p), is maximized. As an item becomes more difficult, it is less likely that <u>any</u> student will score correctly on it. As it becomes less difficult it is more likely that <u>any</u> student will get it correct. This could lead to the suggestion that all items have p=.50, but the usefulness of this suggestion is mitigated by intercorrelations among items. In an extreme case, if the items on a test all intercorrelated perfectly and had difficulties of 0.50, half the examinees would receive a total test grade of zero and the other half would have perfect test scores. Hence, there would be no fine distinctions between examinee's levels of

achievement on whatever trait is being measured. In general, test designers try to choose items with a range of difficulties that average around .50. Items of particularly low difficulty are often included for motivational reasons.

Item discrimination indices were calculated in this study to provide indications that items may be flawed when used with USVI students. Such flaws are ambiguity, the presence of clues, the presence of more than one correct answer, and other technical defects. If none was found upon examinaton of the item, and it was determined that the item did, indeed, appear to measure the objective it was intended to, the item was included in the overall analysis of the results. Any item that discriminates positively can make a contribution to the measurement of pupil achievement and low indices of discrimination are frequently obtained for reasons other than item defects.

Standardized achievement tests are designed to measure several different types of learning outcomes (e.g. knowledge, understanding, application, etc.). Where this is the case, the test items that represent an area receiving relatively little emphasis will tend to have poor discriminating power. For example, if a test has forty items measuring knowledge of specific facts and ten items measuring understanding, the latter items can be expected to have low discriminaton indices. This is because the items measuring understanding have less representation in the total test score and there is typically a low correlation between measures of knowledge and measures of understanding. Low discrimination indices here merely indicate

that these items are measuring something different from what the major part of the test is measuring. Removing such items from the test would make it a more homogeneous measure of knowledge outcomes, but it would also damage the content validity of the test because it would no longer measure objectives in the understanding area. Since achievement test batteries need to measure a wide variety of objectives in a reasonably short period of time, they tend to be fairly heterogeneous in nature and moderately low discrimination indices tend to be the rule rather than the exception.

To summarize, a low discrimination index alerts test users to the posible presence of defects in test items, but does not cause them to discard these items if they appear to be functioning as they should. A well constructed achievement test will, of necessity, contain items with low discriminating power and to discard them would result in a test which is less, rather than more, valid. Due to these considerations, in this study items were examined if they had discrimination indices lower than .20. This is a rather conservative criterion since items that discriminate as low as this may provide useful information, but given the unknown test taking characteristics of USVI students, it was decided to be particularly cautious in the item analysis.

For the most part, items which did not discriminate satisfactorily tended to be those which had extremely high or low difficulty indices (i.e. items which the local sample of students found either very easy or very difficult). In no case did the items seem ambiguous or discriminate negatively.

Student Skills and Test Taking Behaviors

Difficulty indices Difficulty indices for each item on the test were reported out by the test scoring service. In addition, difficulty indices for examinees in the standardization group in the same grade as local examinees at approximately the same time of year were reported. The test scoring service used a chi-squared test for proportions on each difficulty index to test the hypotheses that the proportions of local students scoring correctly on individual items was greater or less than the proportions of examinees in the standardization group scoring correctly at the p=.05 level of significance. Significant differences in either direction were noted.

Level I and Level II Objectives A close look at the difficulty indices tended to disclose a consistent pattern. There appeared to be a set of skills and knowledges which the students in the USVI sample were able to master at levels comparable to students in the standardization sample. From grades 2 through 12, the proportions of students scoring correctly on items testing these skills and knowledges tended to be as high or higher than the proportion of students in the standardization sample. A second set of skills and knowledges seemed to exist which the USVI sample of students appeared to be consistently less successful in mastering than the examinees in the standardization group.

The Taxonomy of Educational Objectives (Bloom, 1956) appears to provide a conceptual hook for understanding these two item groups. The vast majority of items in the first group appear to test objectives which would be classified in the lower three

levels of the taxonomy (i.e. knowledge, comprehension, and application). These include items which require students to spell, compute solutions to mathematical equations using simple algorithms, solve simple, one step mathematical problems to which an algorithm can be directly applied, and to determine explicit meaning in written passages. Most items in the second group appear to test objectives which could be classified in the upper three levels of the taxonomy (i.e. analysis, synthesis, and evaluation). These items require students to solve multistep mathematical word problems, determine relationships, make choices concerning appropriate language useage from context, and determine global, contextual, and inferential meaning from written passages. These findings seem consistent with those of Jensen (1968) concerning the interaction of socioeconomic status and Level I and Level II abilities.

Jensen noted that there were socioeconomic differences in students' abilities to master objectives which would be classified in the three higher levels of the taxonomy (Level II abilities) with non-middle class students having greater difficulty mastering these objectives than students from middle class homes. There were no differences between these two groups of students in their abilities to master objectives in the three lower catagories of the taxonomy (Level I abilities). Table 1 provides a breakdown of the proportions of examinees scoring correctly on items testing Level I and Level II abilitities for

Table 1

Comparisons of Scores on Level I and Level II Objectives

Subtest	LEVEL I OBJECTIVES No. of Items	P Cor. Stan. Grp	P Cor. USVI	LEVEL II OBJECTIVES No. of Items	P Cor. Stan Grp	P Cor USVI
Grade 2						
Vocabulary	37	.67	.69	0	---	---
Reading Comp.	64	.71	.73	23	.71	.72
Word Study Sk.	60	.77	.78	0	---	---
Math Concepts	19	.63	.61	13	.60	.44
Math Comput.	18	.73	.77	14	.64	.54
Listen. Comp.	5	.82	.80	21	.70	.61
Grade 6						
Vocabulary	50	.55	.43	0	---	---
Reading Comp.	23	.63	.59	48	.47	.39
Word Study Sk.	50	.58	.57	0	---	---
Math Concepts	21	.58	.60	15	.50	.36
Math Comput.	45	.54	.56	0	---	---
Math Applic.	18	.62	.59	22	.54	.40
Spelling	60	.55	.59	0	---	---
Language	59	.58	.56	21	.49	.38
Grade 10						
Reading	18	.78	.77	60	.62	.51
English	34	.79	.79	35	.65	.53
Mathematics	21	.75	.72	27	.70	.63

subtests in the batteries given to examinees in grades 2, 6, and 10 as illustrations of this phenomenon. The fact that most USVI public school children come from non-middle class homes while middle class students were represented in the standardization sample tends to support this model as an explanation for these findings.

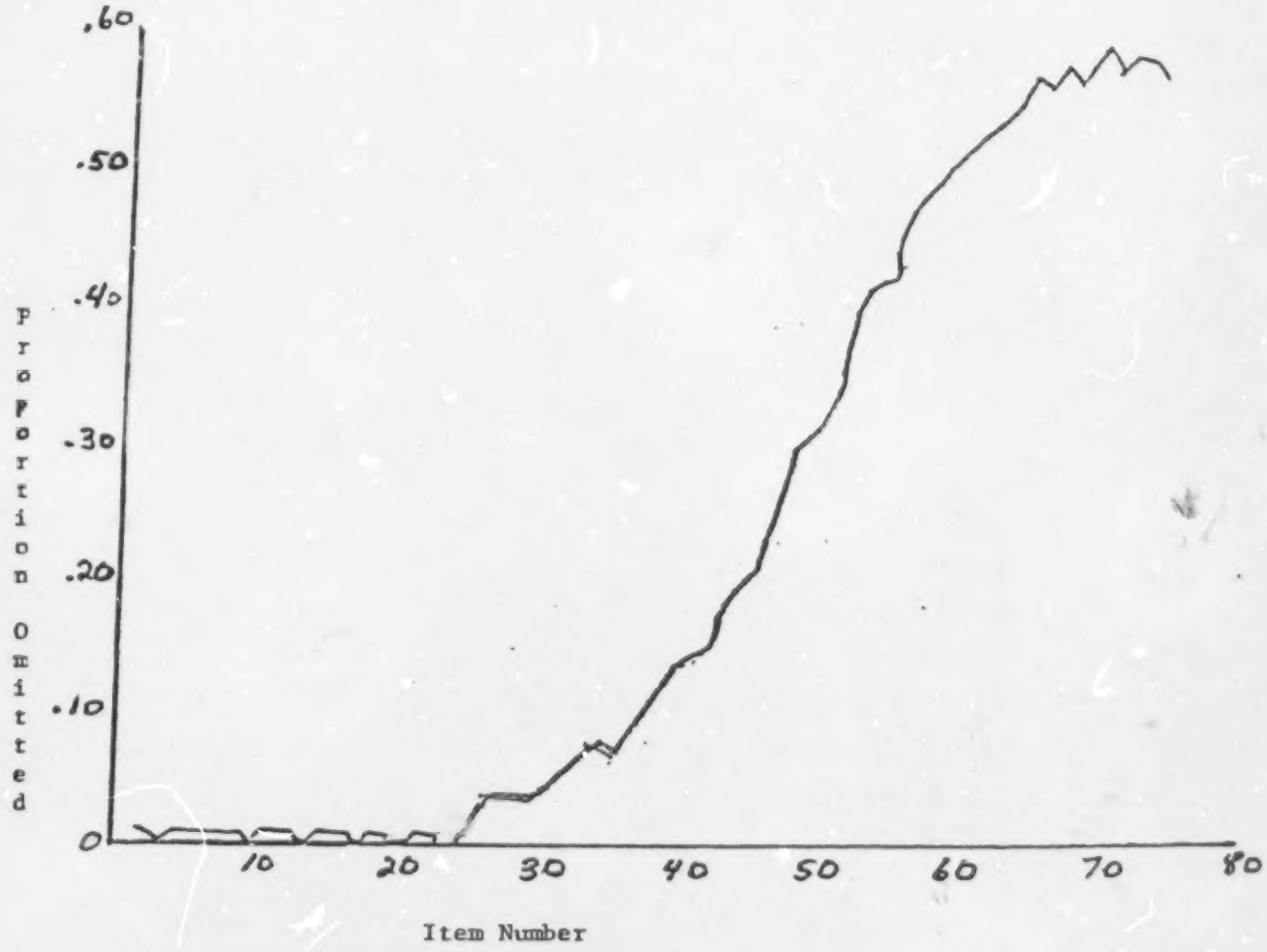
Language In the area of language useage, it was noted that the responses of many students were consistent with the grammar and syntax of the local dialect. This included the dropping of plurals, the confusing of the nomative and objective forms of pronouns with the overuse of the nomative form, the dropping of the indefinite article with the overuse of the definite article, and the dropping of past tenses of verbs. This phenomenon was observed across all grades and is significant since the languge of instruction in the schools is officially standard English and the objectives of the school call for instruction in the use of standard English with absolutely no instruction in the use of dialect.

Omitted responses of the reading test Finally, an examination of the proportions of omitted responses for each item indicated that for all subtests except reading comprehension, examinees had sufficient time to attempt all items when the time recommended by the test publisher was allowed for completion of the subtest. On the reading comprehenson subtest, it was noted that examinees in grades 6 through 12 showed proportions of omits which tended to increase steadily after about the twentieth item on each test with more than 50% of the examinees omitting the last 15 or so items on the tests. Figure 1 shows this phenomenon

Figure 1

Proportions of Omitted Responses on the Reading Comprehension Test

For Grade 8 Examiness



graphically for grade 8 examinees.

A number of explanations for this phenomenon are being considered. The first of these suggests that the students in the local sample are more deliberate readers than their counterparts in the continental United States. They, therefore, read more slowly and do not skim passages to be read. A second is that local examinees have a lower attention span and simply stop taking the test after a certain point and before the end of the testing period. Both of these are rendered plausable by the fact that this phenomenon is not observed in grade 2 where the questions are asked orally by the examiner and the test is divided into two periods, a day apart. Equally noteworthy is the fact that the phenomenon appeared only in the second half of the fourth grade reading test. The first half of the test consisted of a series of independent short answer items while the second consisted of the passage reading type used in the higher grades.

Summation

The model presented and illustrated for the validation of standardized tests for use with students in the English speaking Caribbean appears to be a workable one. It has been demonstrated that the Stanford Achievement Test is appropriate for use in the U.S. Virgin Islands provided that certain modifications are made in the method by which the reading comprehenson tests are administered. Further research is being planned to determine the causes of the difficulties found in these tests at certain grade levels.

Nevertheless, the procedure indicates that the use of these

tests can provide valuable information to teachers and administrators in terms of curriculum effectiveness and cultural differences which might necessitate alterations in test administration procedures which would allow for obtaining valid and reliable scores when these instruments are used.

References

Allen, M.J. & Yen, W.M. Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Co., 1979.

Anastasi, A. Differential Psychology (3rd ed.). New York: Macmillan, 1958.

Bliss, L.B. Validation of the use of the Stanford Achievement Test with U.S.V.I. Students. Virgin Islands of the United States Basic Skills Assessment Survey, Technical Report No. 1. Caribbean Research Institute, College of the Virgin Islands. Paper presented at the Annual Meeting of the American Educational Research Association, New York, 1982. (EDRS Reproduction No. ED 227 150).

Bloom, B. (Ed.) Taxonomy of educational objectives, Handbook I: Cognitive domain. New York: Longmans, Green & Co., 1956.

Chase, C.I. Review of the Test of Academic Skills. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol. 1). Highland Park, NJ: Gryphon Press, 1978.

Deutsch, M. Minority group and class status as related to social and personality factors in scholastic achievement. (Monograph No.2). Ithaca, NY: The Society for Applied Anthropology, 1960.

Ebel, R.L. Review of the 1973 editio of the Stanford Achievment Test. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol. 1). Highland Park, NJ: Gryphon Press, 1978.

Jensen, A.R. Patterns of mental ability and socioeconomic status. Proceedings of the National Academy of Sciences of the United States of America, 1968, 60, 1330-1337

Kasdon, L.M. The Stanford Achievement Test (Review of the 1973 edition of the Stanford Achievement Test). Reading Teacher, 1974, 27, 743+.

Kelly, T.L. The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 1939, 30, 17-24.

Lehmann, I.J. The Stanford Achievement Test Series - 1973 (Review of the 1973 edition of the Stanford Achievement Test). Journal of Educational Measurement, 1975, 12, 297-306.

Thorndike, R.L. Review of the Test of Academic Skills. In O.K. Buros (Ed.), The Eighth Mental Measurements Yearbook (vol. 1). Highland Park, NJ: Gryphon Press, 1978.

Tyler, L. The psychology of individual differences (2nd ed.). New York: Appleton-Century-Crofts, 1956.